**Database Management Systems**

**Project**

**Part III Report**

**Group Member:** Evan Huang, Winston Zhang

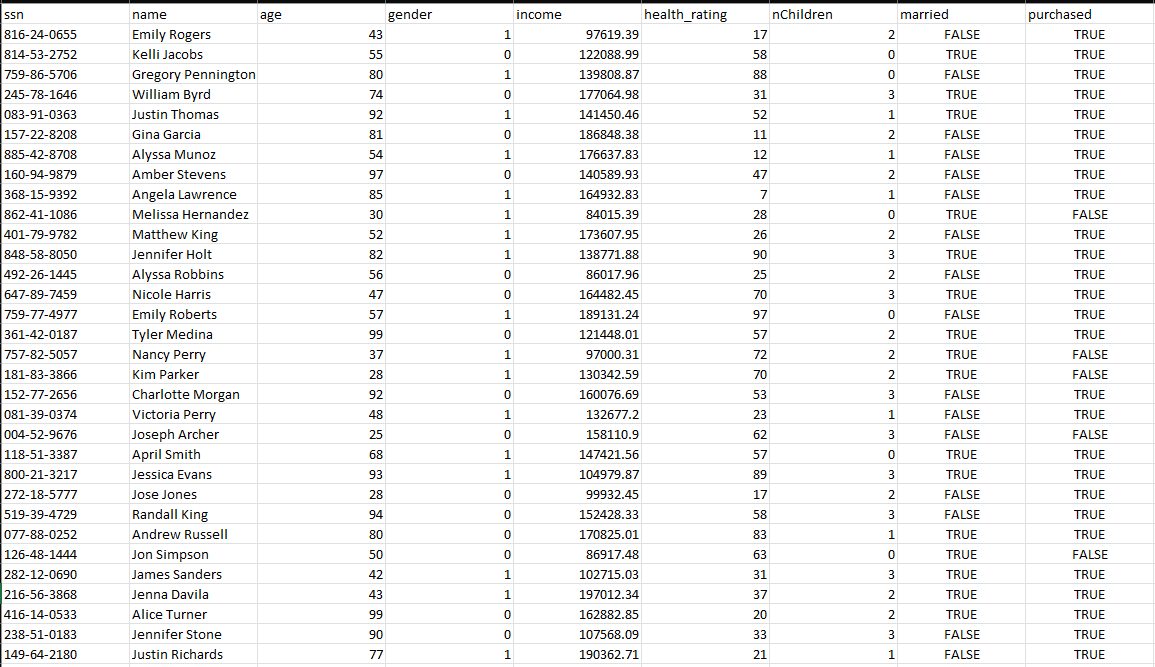
**Database Design:**

Our new database uses AWS RDS MySQL, which now is populated by a table named customer,

with fields such as ssn, name, age, gender, income, health\_rating, nChildren, married, purchased, etc. The reason for the switch is favoring relational and cloud databases, of which we can leverage its computational power to operate data analysis. The database operations are written in main.py.

**Data Pool:**

The data pool, with reference to nationwide statistics and reports, is randomly generated with weights proportional to income, age, health rating, etc. The boolean “purchased” is an evaluation of the combined factors mentioned above, and if it is larger than a certain value(which we use 1.5), then “purchased” is evaluated to TRUE. As an example, the respective purchased percentage for the wealthy group, middle class group, and lower income group for purchasing is 0.76, 0.54, 0.26. Therefore the data pool is to some extent a fit simulation of the real world and is acceptable for machine learning.



*Example data*

**Machine Learning:**

We employed classical machine learning technique K-nearest neighbors on the database to analyze the data. Our goal is to eventually predict the purchase from a customer using the following parameters collected in our database: customer’s age, gender, income, health rating, how many children he/she has and marital status. Once this objective is achieved, the company will have a better chance of spotting the target crowd for its products, therefore boosting its performance.

We chose the data based on the following reasons: Age is a primary factor for customers to consider purchasing additional insurance products and a key factor for the company to assess risk. Higher income levels might correlate with a greater chance to invest in insurance products based on simple expected income effects. Health rating is the most important factor for determining both risks borne by the company and the likelihood of purchase. The number of children is likely to determine the purchase probability of certain types of insurances such as life insurance. Marital status is likely to alter the need for insurance.

KNN algorithm is chosen due to its effectiveness and suitability for our dataset and objectives. KNN is a non-parametric algorithm, meaning it makes no assumptions about the underlying data distribution which is desirable in this case. Moreover, KNN is a suitable algorithm for classification tasks, especially when both numerical(age, children, health rating) and categorical(marital status, gender, purchased) are considered.

We implemented KNN algorithm in python(ml.py):

* Connect to the database and read/preprocess the data using SQL query
* Splitting the data into usual 80/20 training and testing datasets
* Training the KNN model on the training set and test the model on the testing dataset
* Assess the performance of the model using confusion matrix and classification report